

Computationally Tractable Algorithms for Finding a Subset of Non-defective Items from a Large Population

Abhay Sharma and Chandra R. Murthy

Dept. of ECE, Indian Institute of Science, Bangalore 560 012, India

abhay.bits@gmail.com, cmurthy@ece.iisc.ernet.in

Abstract

In the classical non-adaptive group testing setup, pools of items are tested together, and the main goal of a recovery algorithm is to identify the *complete defective set* given the outcomes of different group tests. In contrast, the main goal of a *non-defective subset recovery* algorithm is to identify a *subset* of non-defective items given the test outcomes. In this paper, we present a suite of computationally efficient and analytically tractable non-defective subset recovery algorithms. By analyzing the probability of error of the algorithms, we obtain bounds on the number of tests required for non-defective subset recovery with arbitrarily small probability of error. Our analysis accounts for the impact of both the additive noise (false positives) and dilution noise (false negatives). By comparing with the information theoretic lower bounds, we show that the upper bounds on the number of tests are order-wise tight up to a $\log^2 K$ factor, where K is the number of defective items. We also provide simulation results that compare the relative performance of the different algorithms and provide further insights into their practical utility. The proposed algorithms significantly outperform the straightforward approaches of testing items one-by-one, and of first identifying the defective set and then choosing the non-defective items from the complement set, in terms of the number of measurements required to ensure a given success rate.

Index Terms

Non-adaptive group testing, boolean compressed sensing, non-defective subset recovery, inactive subset identification, linear program analysis, combinatorial matching pursuit, sparse signal models.

I. INTRODUCTION

The general group testing framework [1], [2] considers a large set of N items, in which an unknown subset of K items possess a certain testable property, e.g., the presence of an antigen in a blood sample, presence of a pollutant in an air sample, etc. This subset is referred to as the “defective” subset, and its complement is referred to as the “non-defective” or “healthy” subset. A defining notion of this framework is the *group test*, a test that operates on a *group* of items and provides a binary indication as to whether or not the property of interest is present collectively in the group. A *negative* indication implies that none of the tested items are defective. A *positive* indication implies that at least one of the items is defective. In practice, due to the hardware and test procedure limitations, the group tests are not completely reliable. Using the outcomes of multiple such (noisy) group tests, a basic goal of group testing is to reliably identify the defective set of items with as few tests as possible. The framework of group testing has found applications in diverse engineering fields such as industrial testing [3], DNA sequencing [2], [4], data pattern mining [5]–[7], medical screening [2], multi-access communications [2], [8], data streaming [9], [10], etc.

One of the popular versions of the above theme is the non-adaptive group testing (NGT), where different tests are conducted simultaneously, i.e., the tests do not use information provided by the outcome of any other test. NGT is especially useful when the individual tests are time consuming, and hence the testing time associated with adaptive, sequential testing is prohibitive. An important aspect of NGT is how to determine the set of individuals that go into each group test. Two main approaches exist: a combinatorial approach, see e.g., [11]–[13], which considers explicit constructions of test matrices/pools; and a probabilistic pooling approach, see e.g., [10], [14], [15], where the items included in the group test are chosen uniformly at random from the population. Non-adaptive group testing has also been referred to as *boolean* compressed sensing in the recent literature [16], [17].

In this work, in contrast to the defective set identification problem, we study the *healthy/non-defective subset identification* problem in the noisy, non-adaptive group testing with random pooling (NNGT-R) framework. There are many applications where the goal is to identify only a small subset of non-defective items. For example, consider the spectrum hole search problem in a cognitive radio (CR) network setup. It is known that the primary user occupancy is sparse in the frequency domain, over a wide band of

interest [18], [19]. This is equivalent to having a small subset of defective items embedded in a large set of candidate frequency bins. The secondary users do not need to identify all the frequency bins occupied by the primary users; they only need to discover a small number of unoccupied sub-bands to setup the secondary communications. This, in turn, is a non-defective subset identification problem when the bins to be tested for primary occupancy can be pooled together into group tests [20]. In [21], using information theoretic arguments, it was shown that compared to the conventional approach of identifying the non-defective subset by first identifying the defective set, directly searching for an L -sized non-defective subset offers a reduction in the number of tests, especially when L is small compared to $N - K$. The achievability results in [21] were obtained by analyzing the performance of the exhaustive search based algorithms which are not practically implementable. In this paper, we develop computationally efficient algorithms for non-defective subset identification in an NNGT-R framework.

We note that the problem of non-defective subset identification is a generalization of the defective set identification problem, in the sense that, when $L = N - K$, the non-defective subset identification problem is identical to that of identifying the K defective items. Hence, by setting $L = N - K$, the algorithms presented in this work can be related to algorithms for finding the defective set; see [2] for an excellent collection of existing results and references. In general, for the NNGT-R framework, three broad approaches have been adopted for defective set recovery [17]. First, the row based approach (also frequently referred to as the “naïve” decoding algorithm) finds the defective set by finding *all* the non-defective items. The survey in [22] lists many variants of this algorithm for finding defective items. More recently, the **CoCo** algorithm was studied in [17], where an interesting connection of the naïve decoding algorithm with the classical coupon-collector problem was established for the noiseless case. The second popular decoding approach is based on the idea of finding defective items iteratively (or greedily) by matching the column of the test matrix corresponding to a given item with the test outcome vector [2], [17], [23], [24]. For example, in [23], column matching consists of taking set differences between the set of pools where the item is tested and the set of pools with positive outcomes. Another variant of matching is considered in [17], where, for a given column, the ratio of number of times an item is tested in pools with positive and negative outcomes is computed and compared to a threshold. A recent work, [25], investigates the problem of finding zeros in a sparse vector in the compressive sensing framework, and also proposes a greedy algorithm based on correlating the columns of the sensing matrix (i.e., column

matching) with the output vector.¹ The connection between defective set identification in group testing and the sparse recovery in compressive sensing was further highlighted in [7], [17], [26], where relaxation based linear programming algorithms have been proposed for defective set identification in group testing. A class of linear programs to solve the defective set identification problem was proposed by letting the boolean variables take real values (between 0 and 1) and setting up inequality or equality constraints to model the outcome of each pool.

In this work, we develop novel algorithms for identifying a non-defective subset in an NNGT-R framework. We present error rate analysis for each algorithm and derive non-asymptotic upper bounds on the average error rate. The derivation leads to a theoretical guarantee on the sample complexity, i.e., the number of tests required to identify a subset of non-defective items with arbitrarily small probability of error. We summarize our main contributions as follows:

- We propose a suite of computationally efficient and analytically tractable algorithms for identifying a non-defective subset of given size in a NNGT-R framework: **RoAI** (row based), **CoAI** (column based) and **RoLpAI**, **RoLpAI++**, **CoLpAI** (Linear Program (LP) relaxation based) algorithms.
- We derive bounds on the number of tests that guarantee successful non-defective subset recovery for each algorithm. The derived bounds are a function of the system parameters, namely, the number of defective items, the size of non-defective subset, the population size, and the noise parameters. Further,
 - The presented bounds on the number of tests for different algorithms are within $O(\log^2 K)$ factor, where K is the number of defective items, of the information theoretic lower bounds which were derived in our past work [21].
 - For our suite of LP based algorithms, we present a novel analysis technique based on characterizing the recovery conditions via the dual variables associated with the LP, which may be of interest in its own right.
- Finally, we present numerical simulations to compare the relative performance of the algorithms.

¹Note that directly computing correlations between column vector for an item and the test outcome vector will not work in case of group testing, as both the vectors are boolean. Furthermore, positive and negative pools have asymmetric roles in the group testing problem.

The results also illustrate the significant benefit in finding non-defective items directly, compared to using the existing defective set recovery methods or testing items one-by-one, in terms of the number of group tests required.

The rest of the paper is organized as follows. Section II describes the NNGT-R framework and the problem setup. The proposed algorithms and the main analytical results are presented in Section III. The proofs of the main results are provided in Section V. Section VI discusses the numerical simulation results, and the conclusions are presented in Section VII. We conclude this section by presenting the notation followed throughout the paper.

Notation: Matrices are denoted using uppercase bold letters and vectors are denoted using an underline. For a given matrix \mathbf{A} , $\underline{a}_i^{(r)}$ and \underline{a}_i denote the i^{th} row and column, respectively. For a given index set S , $\mathbf{A}(S, :)$ denotes a sub-matrix of \mathbf{A} where only the rows indexed by set S are considered. Similarly, $\mathbf{A}(:, S)$ or \mathbf{A}_S denotes a sub-matrix of \mathbf{A} that consists only of columns indexed by set S . For a vector \underline{a} , $\underline{a}(i)$ denotes its i^{th} component; $\text{supp}(\underline{a}) \triangleq \{j : \underline{a}(j) > 0\}$; $\{\underline{a} = c\}$ denotes the set $\{j : \underline{a}(j) = c\}$ for any c . In the context of a boolean vector, \underline{a}^c denotes the component wise boolean complement of \underline{a} . $\underline{1}_n$ and $\underline{0}_n$ denote an all-one and all-zero vector, respectively, of size $n \times 1$. We denote the component wise inequality as $\underline{a} \preceq \underline{b}$, i.e., it means $\underline{a}(i) \leq \underline{b}(i) \forall i$. Also, $\underline{a} \circ \underline{b}$ denotes the component-wise product, i.e., $(\underline{a} \circ \underline{b})(i) = \underline{a}(i)\underline{b}(i), \forall i$. The boolean OR operation is denoted by “ \vee ”. For any $q \in [0, 1]$, $\mathcal{B}(q)$ denotes the Bernoulli distribution with parameter q . $\mathbb{I}_{\mathcal{A}}$ denotes the indicator function and returns 1 if the event \mathcal{A} is true, else returns 0. Note that, $x(n) = O(y(n))$ implies that $\exists B > 0$ and $n_0 > 0$, such that $|x(n)| \leq B|y(n)|$ for all $n > n_0$. Further, $x(n) = \Omega(y(n))$ implies that $\exists B > 0$ and $n_0 > 0$, such that $|x(n)| \geq B|y(n)|$ for all $n > n_0$. Also, $x(n) = o(y(n))$ implies that for every $\epsilon > 0$, there exists an $n_0 > 0$ such that $|x(n)| \leq \epsilon|y(n)|$ for all $n > n_0$. All logarithms in this papers are to the base e . Also, for any $p \in [0, 1]$, $H_b(p)$ denotes the binary entropy in nats, i.e., $H_b(p) \triangleq -p \log(p) - (1-p) \log(1-p)$.

II. SIGNAL MODEL

In our setup, we have a population of N items, out of which K are defective. Let $\mathcal{G} \subset [N]$ denote the defective set, such that $|\mathcal{G}| = K$. We consider a non-adaptive group testing framework with random pooling [2], [16], [17], [27], where the items to be pooled in a given test are chosen at random from the population. The group tests are defined by a boolean matrix, $\mathbf{X} \in \{0, 1\}^{M \times N}$, that assigns different items

to the M group tests (pools). The j^{th} pool tests the items corresponding to the columns with 1 in the j^{th} row of \mathbf{X} . We consider an i.i.d. random Bernoulli measurement matrix [16], where each $X_{ij} \sim \mathcal{B}(p)$ for some $0 < p < 1$. Thus, M randomly generated pools are specified. In the above, p is a design parameter that controls the average group size, i.e., the average number of items being tested in a single group test. In particular, we choose $p = \frac{\alpha}{K}$, and a specific value of α is chosen based on the analysis of different algorithms.

If the tests are completely reliable, then the output of the M tests is given by the boolean OR of the columns of \mathbf{X} corresponding to the defective set \mathcal{G} . However, in practice, the outcome of a group test may be unreliable. Two popular noise models that are considered in the literature on group testing are [16], [17], [23]: (a) An *additive* noise model, where there is a probability, $q \in (0, 0.5)$, that the outcome of a group test containing only non-defective items turns out to be positive (Fig. 1); (b) A *dilution* model, where there is a probability, $u \in (0, 0.5)$, that a given item does not participate in a given group test (see Fig. 1). Let $\underline{d}_i \in \{0, 1\}^M$. Let $\underline{d}_i(j) \sim \mathcal{B}(1 - u)$ be chosen independently for all $j = 1, 2, \dots, M$ and for all $i = 1, 2, \dots, N$. Let $\mathbf{D}_i \triangleq \text{diag}(\underline{d}_i)$. The output vector $\underline{y} \in \{0, 1\}^M$ can be represented as

$$\underline{y} = \bigvee_{i=1}^N \mathbf{D}_i \underline{x}_i \mathbb{I}_{\{i \in \mathcal{G}\}} \bigvee \underline{w}, \quad (1)$$

where $\underline{x}_i \in \{0, 1\}^M$ is the i^{th} column of \mathbf{X} , $\underline{w} \in \{0, 1\}^M$ is the additive noise with the i^{th} component $w(i) \sim \mathcal{B}(q)$. Note that, for the noiseless case, $u = 0, q = 0$. Given the test output vector, \underline{y} , our goals are as follows:

- (a) To find computationally tractable algorithms to identify L non-defective items, i.e., an L -sized subset belonging to $[N] \setminus \mathcal{G}$.
- (b) To analyze the performance of the proposed algorithms with the objective of (i) finding the number of tests and (ii) choosing the appropriate design parameters that leads to non-defective subset recovery with high probability of success.

In the literature on defective set recovery in group testing or on sparse vector recovery in compressed sensing, there exist two type of recovery results: (a) *Non-uniform/Per-Instance recovery results*: These state that a randomly chosen test matrix leads to non-defective subset recovery with high probability of success for a given fixed defective set and, (b) *Uniform/Universal recovery results*: These state that a random draw of the test matrix leads to a successful non-defective subset recovery with high probability

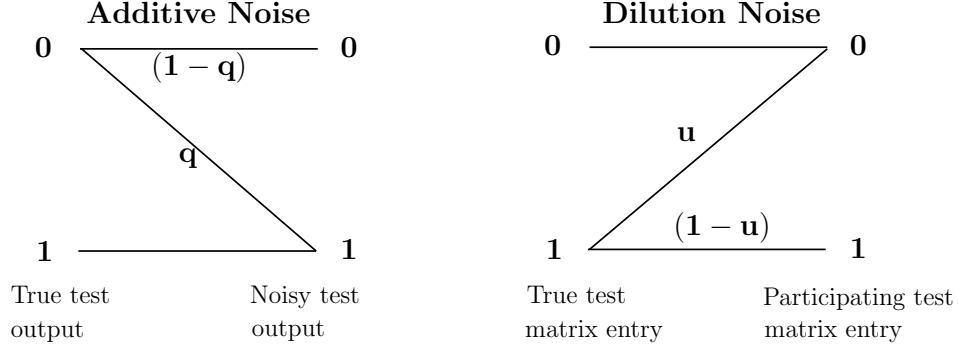


Fig. 1. Impact of different types of noise on the group testing signal model.

for all possible defective sets. It is possible to easily extend non-uniform results to the uniform case using union bounds. Hence, we focus mainly on non-uniform recovery results, and demonstrate the extension to the uniform case for one of the proposed algorithms (see Corollary 1). Note that the non-uniform scenario is equivalent to the uniform recovery scenario when the defective set is chosen uniformly at random from the set of $\binom{N}{K}$ possible choices. For the latter scenario, information theoretic lower bounds on the number of tests for the non-defective subset recovery problem were derived in [21] using Fano's inequality. We use these bounds in assessing the performance of the proposed algorithms (see Section IV). For the ease of reference, we summarize these results in Table I.

For later use, we summarize some key facts pertaining to the above signal model in the lemma below. For any $l \in [M]$ and $k \in [N]$, let X_{lk} denote the $(l, k)^{\text{th}}$ entry of the test matrix \mathbf{X} and let $Y_l \triangleq \underline{y}(l)$ denote the l^{th} test output. With u , q and p as defined above, let $\Gamma \triangleq (1 - q)(1 - (1 - u)p)^K$ and $\gamma_0 \triangleq \frac{u}{(1 - (1 - u)p)}$. Then it follows that,

Lemma 1. (a) $\mathbb{P}(Y_l = 0) = \Gamma$.

(b) For any $j \notin S_d$, $\mathbb{P}(Y_l | X_{lj}) = \mathbb{P}(Y_l)$.

(c) For any $i \in S_d$, $\mathbb{P}(Y_l = 0 | X_{li} = 1) = \gamma_0 \Gamma$ and $\mathbb{P}(Y_l = 0 | X_{li} = 0) = \frac{\Gamma}{1 - (1 - u)p}$. Further, using Bayes rule, $\mathbb{P}(X_{li} = 1 | Y_l = 0) = p\gamma_0$.

(d) Given Y_l , X_{li} is independent of X_{lj} for any $i \in S_d$ and $j \notin S_d$.

The proof is provided in Appendix A.

III. ALGORITHMS AND MAIN RESULTS

We now present several algorithms for non-defective/healthy subset recovery. Each algorithm takes the observed noisy test-output vector $\underline{y} \in \{0, 1\}^M$ and the test matrix $\mathbf{X} \in \{0, 1\}^{M \times N}$ as inputs, and outputs

a set of L items, \hat{S}_L , that have been declared non-defective. The recovery is successful if the declared set does not contain any defective item, i.e., $\hat{S}_L \cap S_d = \{\emptyset\}$. For each algorithm, we derive expressions for the upper bounds on the average probability of error, which are further used in deriving the number of tests required for successful non-defective subset recovery.

A. Row Based Algorithm

Our first algorithm to find non-defective items is also the simplest and the most intuitive one. We make use of the basic fact of group testing that, in the noiseless case, if the test outcome is negative, then all the items being tested are non-defective.

RoAI (Row based algorithm):

- Compute $\underline{z} = \sum_{j \in \text{supp}(\underline{y}^c)} \underline{x}_j^{(r)}$, where $\underline{x}_j^{(r)}$ is the j^{th} row of the test matrix.
- Order entries of \underline{z} in descending order.
- Declare the items indexed by the top L entries as the non-defective subset.

That is, declare the L items that have been tested most number of times in pools with negative outcomes as non-defective items. The above decoding algorithm proceeds by only considering the tests with negative outcomes. Note that, when the test outcomes are noisy, there is a nonzero probability of declaring a defective item as non-defective. In particular, the dilution noise can lead to a test containing defective items in the pool being declared negative, leading to a possible misclassification of the defective items. On the other hand, since the algorithm only considers tests with negative outcomes, additive noise does not lead to misclassification of defective items as non-defective. However, the additive noise does lead to an increased number of tests as the algorithm has to possibly discard many of the pools that contain only non-defective items.

We note that existing row based algorithms for finding defective set [2], [17] can be obtained as a special case of the above algorithm by setting $L = N - K$, i.e., by looking for all non-defective items. However, the analysis in the past work does not quantify the impact of the parameter L and that is our main goal here. We characterize the number of tests, M , that are required to find L non-defective items with high probability of success using **RoAI** in Theorem 1.

B. Column Based Algorithm

The column based algorithm is based on matching the columns of the test matrix with the test outcome vector. A non-defective item does not impact the output and hence the corresponding column in the test matrix should be “uncorrelated” with the output. On the other hand, “most” of the pools that test a defective item should test positive. This forms the basis of distinguishing a defective item from a non-defective one. The specific algorithm is as follows:

CoAI (Column based algorithm): Let $\psi_{cb} \geq 0$ be any constant.

- For each $i = 1, \dots, N$, compute

$$\mathcal{T}(i) = \underline{x}_i^T \underline{y}^c - \psi_{cb}(\underline{x}_i^T \underline{y}), \quad (2)$$

where \underline{x}_i is the i^{th} column of \mathbf{X} .

- Sort $\mathcal{T}(i)$ in descending order.
- Declare the items indexed by the top L entries as the non-defective subset.

We note that, in contrast to the row based algorithm, **CoAI** works with pools of both the negative and positive test outcomes (when the parameter $\psi_{cb} > 0$; its choice is explained below). For both **RoAI** and **CoAI**, by analyzing the probability of error, we can derive the sufficient number of tests required to achieve arbitrarily small error rates. We summarize the main result in the following theorem:

Theorem 1. (Non-Uniform recovery with **RoAI** and **CoAI**) Let $\Gamma \triangleq (1 - q)(1 - (1 - u)p)^K$ and $\gamma_0 \triangleq \frac{u}{(1 - (1 - u)p)}$. Suppose $K > 1$ and let p be chosen as $\frac{\alpha}{K}$ with $\alpha = \frac{1}{(1 - u)}$. For **RoAI**, let $\psi_0 \triangleq 0$. For **CoAI**, choose $\psi_0 \triangleq \frac{\gamma_0 \Gamma}{1 - \gamma_0 \Gamma}$ and set $\psi_{cb} = \psi_0$. Let $c_0 > 0$ be any constant. Then, there exist absolute constants $C_{a1}, C_{a2} > 0$ independent of N, L and K , and different for each algorithm, such that, if the number of tests is chosen as

$$M \geq (1 + c_0) \frac{K(1 - u)}{(1 - q)(1 - \gamma_0)^2(1 + \psi_0)} \left(\frac{C_{a1} \log \left[K \binom{N-K}{L-1} \right]}{(N - K) - (L - 1)} + C_{a2} \log K \right), \quad (3)$$

then, for a given defective set, the algorithms **RoAI** and **CoAI** find L non-defective items with probability exceeding $1 - \exp \left(-c_0 \log \left(K \binom{N-K}{L-1} \right) \right) - \exp(-c_0 \log K)$.

The following corollary extends Theorem 1 to uniform recovery of a non-defective subset using **RoAI** and **CoAI**.

Corollary 1. (Uniform recovery with **RoAI** and **CoAI**) For any positive constant $c_0 > 0$, there exist

absolute constants $C_{a1}, C_{a2} > 0$ independent of N, L and K , and different for each algorithm, such that if the number of tests is chosen as

$$M \geq (1 + c_0) \frac{K(1 - u)}{(1 - q)(1 - \gamma_0)^2(1 + \psi_0)} \left(\frac{C_{a1} \log \left[K \binom{N-K}{L-1} \binom{N}{K} \right]}{(N - K) - (L - 1)} + C_{a2} \log N \right), \quad (4)$$

then for **any** defective set, the algorithms **RoAI** and **CoAI** find L non-defective items with probability exceeding $1 - \exp \left(-c_0 \log \left(K \binom{N-K}{L-1} \right) \right) - \exp(-c_0 \log N)$.

The proof of the above theorem is presented in Section V-A. It is tempting to compare the performance of **RoAI** and **CoAI** by comparing the required number of tests as presented in (3). However, such comparisons must be done keeping in mind that the required number of observations in (3) are based on an upper bound on the average probability of error. The main objective of these results is to provide a guarantee on the number of tests required for non-defective subset recovery and highlight the order-wise dependence of the number of tests on the system parameters. For the comparison of the relative performance of the algorithms, we refer the reader to Section VI, where we present numerical results obtained from simulations. From the simulations, we observe that **CoAI** performs better than **RoAI** for most scenarios of interest. This is because, in contrast to **RoAI**, **CoAI** uses the information obtained from pools corresponding to both negative and positive test outcomes.

C. Linear program relaxation based algorithms

In this section, we consider linear program (LP) relaxations to the non-defective subset recovery problem and identify the conditions under which such LP relaxations lead to recovery of a non-defective subset with high probability of success. These algorithms are inspired by analogous algorithms studied in the context of defective set recovery in the literature [17], [26]. However, past analysis on the number of tests for the defective set recovery do not carry over to the non-defective subset recovery because the goals of the algorithms are very different. Let $Y_z \triangleq \{l \in [M] : \underline{\mathbf{y}}(l) = 0\}$, i.e., Y_z is the index set of all the pools whose test outcomes are negative and $M_z \triangleq |Y_z|$. Similarly, let $Y_p \triangleq \{l \in [M] : \underline{\mathbf{y}}(l) = 1\}$ and

$M_p \triangleq |Y_p|$. Define the following linear program, with optimization variables $\underline{z} \in \mathbb{R}^N$ and $\underline{\eta}_z \in \mathbb{R}^{M_z}$:

$$\underset{\underline{z}, \underline{\eta}_z}{\text{minimize}} \quad \underline{1}_{M_z}^T \underline{\eta}_z \quad (5)$$

$$\text{(LP0)} \quad \text{subject to} \quad \mathbf{X}(Y_z, :)(\underline{1}_N - \underline{z}) - \underline{\eta}_z = \underline{0}_{M_z}, \quad (6)$$

$$\underline{0}_N \preceq \underline{z} \preceq \underline{1}_N, \quad \underline{\eta}_z \succeq \underline{0}_{M_z},$$

$$\underline{1}_N^T \underline{z} \leq L.$$

Consider the following algorithm:²

RoLpAl (LP relaxation with negative outcome pools only)

- Setup and solve **LP0**. Let $\hat{\underline{z}}$ be the solution of **LP0**.
- Sort $\hat{\underline{z}}$ in descending order.
- Declare the items indexed by the top L entries as the non-defective subset.

The above program relaxes the combinatorial problem of choosing L out of N items by allowing the boolean variables to acquire “real” values between 0 and 1 as long as the constraints imposed by negative pools, specified in (6), are met. Intuitively, the variable \underline{z} (or the variable $[\underline{1}_N - \underline{z}]$) can be thought of as the confidence with which an item is being declared as non-defective (or defective). The constraint $\underline{1}_N^T \underline{z} \leq L$ forces the program to assign high values (close to 1) for “approximately” the top L entries only, which are then declared as non-defective.

For the purpose of analysis, we first derive sufficient conditions for correct non-defective subset recovery with **RoLpAl** in terms of the dual variables of **LP0**. We then derive the number of tests required to satisfy these sufficiency conditions with high probability. The following theorem summarizes the performance of the above algorithm:

Theorem 2. (Non-Uniform recovery with **RoLpAl**) Let $K > 1$ and let p be chosen as $\frac{\alpha}{K}$ with $\alpha = \frac{1}{(1-u)}$. If the number of tests is chosen as in (3) with $\psi_0 = 0$, then for a given defective set there exist absolute constants $C_{a1}, C_{a2} > 0$ independent of N, L and K , such that **RoLpAl** finds L non-defective items with probability exceeding $1 - \exp\left(-c_0 \log\left(K \binom{N-K}{L-1}\right)\right) - \exp(-c_0 \log K)$.

²The other algorithms presented in this sub-section, namely, **RoLpAl++** and **CoLpAl**, have the same structure and differ only in the linear program being solved.

The proof of the above theorem is presented in Section V-B. Note that **LP0** operates only on the set of pools with negative outcomes and is, thus, sensitive to the dilution noise which can lead to a misclassification of a defective item as non-defective. To combat this, we can leverage the information available from the pools with positive outcomes also, by incorporating constraints for variables involved in these tests. Consider the following linear program with optimization variables $\underline{z} \in \mathbb{R}^N$ and $\underline{\eta}_z \in \mathbb{R}^{M_z}$:

$$\begin{aligned}
 & \underset{\underline{z}, \underline{\eta}_z}{\text{minimize}} && \underline{1}_{M_z}^T \underline{\eta}_z && (7) \\
 (\text{LP1}) \quad & \text{subject to} && \mathbf{X}(Y_z, :)(\underline{1}_N - \underline{z}) - \underline{\eta}_z = \underline{0}_{M_z} \\
 & && \mathbf{X}(Y_p, :)(\underline{1}_N - \underline{z}) \succcurlyeq (1 - \epsilon_0)\underline{1}_{M_p} && (8) \\
 & && \underline{0}_N \preccurlyeq \underline{z} \preccurlyeq \underline{1}_N, \quad \underline{\eta}_z \succcurlyeq \underline{0}_{M_z} \\
 & && \underline{1}_N^T \underline{z} \leq L.
 \end{aligned}$$

In the above, $0 < \epsilon_0 \ll 1$ is a small positive constant. Note that (8) attempts to model, in terms of real variables, a boolean statement that at least one of the items tested in tests with positive outcomes is a defective item. We refer to the algorithm based on **LP1** as **RoLpAl++**. We expect **RoLpAl++** to outperform **RoLpAl**, as the constraint (8) can provide further differentiation between items that are indistinguishable just on the basis of negative pools. Note that, due to the constraint $\underline{1}_N^T \underline{z} \leq L$, the entries of $\hat{\underline{z}}$ in $[N] \setminus \hat{S}_L$ are generally assigned small values. Hence, when L is small, for many of the positive pools, the constraint (8) may not be active. Thus, we expect **RoLpAl++** to perform better than **RoLpAl** as the value of L increases; this will be confirmed via simulation results in Section VI. Due to the difficulty in obtaining estimates for the dual variables associated with the constraints (8), it is difficult to derive theoretical guarantees for **RoLpAl++**. However, we expect the guarantees for **RoLpAl++** to be similar to **RoLpAl**, and we refer the reader to Appendix F for a discussion regarding the same.

Motivated by the connection between **RoAl** and **RoLpAl**, as revealed in the proof of Theorem 2 (see Section V-B), we now propose another LP based non-defective subset recovery algorithm that incorporates both positive and negative pools, which, in contrast to **RoLpAl++**, turns out to be analytically tractable. By incorporating (8) in an unconstrained form and by using the *same* weights for all the associated

Lagrangian multipliers in the optimization function, we get:

$$\begin{aligned}
 & \underset{\underline{z}}{\text{minimize}} && \underline{1}_{M_z}^T \mathbf{X}(Y_z, :)(\underline{1}_N - \underline{z}) - \psi_{lp} \left[\underline{1}_{M_p}^T \mathbf{X}(Y_p, :)(\underline{1}_N - \underline{z}) \right] \\
 (\mathbf{LP2}) \quad & \text{subject to} && \underline{0}_N \preceq \underline{z} \preceq \underline{1}_N, \\
 & && \underline{1}_N^T \underline{z} \leq L,
 \end{aligned} \tag{9}$$

where $\psi_{lp} > 0$ is a positive constant that provides appropriate weights to the two different type of cumulative errors. Note that, compared to **LP1**, we have also eliminated the equality constraints in the above program. The basic intuition is that by using (8) in an unconstrained form, i.e., by maximizing $\sum_{j \in Y_z} \mathbf{X}(j, :)(\underline{1}_N - \underline{z})$, the program will tend to assign higher values to $(1 - \hat{z}(i))$ (and hence lower values to $\hat{z}(i)$) for $i \in S_d$ since for random test matrices with i.i.d. entries, the defective items are likely to be tested more number of times in the pools with positive outcomes. Also, in contrast to **LP1** where different weightage is given to each positive pool via the value of the associated dual variable, **LP2** gives the same weightage to each positive pool, but it adjusts the overall weightage of positive pools using the constant ψ_{lp} . We refer to the algorithm based on **LP2** as **CoLpAl**. The theoretical analysis for **CoLpAl** follows on similar lines as **RoLpAl** and we summarize the main result in the following theorem:

Theorem 3. (Non-Uniform recovery with **CoLpAl**) Let $\Gamma \triangleq (1-q)(1-(1-u)p)^K$ and $\gamma_0 \triangleq \frac{u}{(1-(1-u)p)}$. Let $K > 1$ and let p be chosen as $\frac{\alpha}{K}$ with $\alpha = \frac{1}{(1-u)}$. Let $\psi'_0 \triangleq \min \left(\frac{\gamma_0 \Gamma}{1-\gamma_0 \Gamma}, \frac{\Gamma}{2(1-\Gamma)} \right)$ and set $\psi_{lp} = \psi'_0$. Then, for any positive constant c_0 , there exist absolute constants $C_{a1}, C_{a2} > 0$ independent of N, L and K , such that, if the number of tests is chosen as in (3) with $\psi_0 = 0$, then for a given defective set **CoLpAl** finds L non-defective items with probability exceeding $1 - 2 \exp \left(-c_0 \log \left(K \binom{N-K}{L-1} \right) \right) - \exp(-c_0 \log K)$.

An outline of the proof of the above theorem is presented in Section V-C.

IV. DISCUSSION ON THE THEORETICAL GUARANTEES

We now present some interesting insights by analyzing the number of tests required for correct non-defective subset identification by the proposed recovery algorithms. We note that the expression in (3) adapted for different algorithms differs only on account of the constants involved. This allows us to present a unified analysis for all the algorithms.

- (a) Asymptotic analysis of M as $N \rightarrow \infty$: We consider the parameter regimes where $K, L \rightarrow \infty$ as $N \rightarrow \infty$. We note that, under these regimes, when the conditions specified in the theorems

are satisfied, the probability of decoding error can be made arbitrarily close to zero. In particular, we consider the regime where $\frac{K}{N} \rightarrow \beta_0$, $\frac{L}{N} \rightarrow \alpha_0$, as $N \rightarrow \infty$, where $0 \leq \beta_0 < \alpha_0 < 1$, $\alpha_0 + \beta_0 < 1$. Define $\zeta \triangleq \frac{L-1}{N-K}$, and $\zeta \rightarrow \zeta_0 \triangleq \frac{\alpha_0}{1-\beta_0}$ as $N \rightarrow \infty$. Also, note that $\gamma_0 \rightarrow u$ as $N \rightarrow \infty$. Using Stirling's formula, it can be shown that $\lim_{N \rightarrow \infty} \frac{\log \binom{N-K}{L-1}}{(N-K)-(L-1)} \leq \frac{H_b(\zeta_0)}{1-\zeta_0}$ (see, [21]), where $H_b(\cdot)$ is the binary entropy function. Further, let $g(\zeta) \triangleq \frac{H_b(\zeta)}{1-\zeta}$. Now, since $g(\zeta_0)$ is a constant, the sufficient number of tests M for the proposed algorithms depends on K as $M \geq C_0 \frac{K}{(1-u)(1-q)} (C_{a1}g(\zeta_0) + C_{a2} \log K + o(1))$. Here, C_0, C_{a1} and C_{a2} are constants independent of N, K, L, u and q .

We compare the above with the sufficient number of test required for the defective set recovery algorithms. When K grows sub-linearly with N (i.e., $\beta_0 = 0$), the sufficient number of tests for the proposed decoding algorithms is $O(K \log K)$, which is better than the sufficient number of tests for finding the defective set, which scales as $O(K \log N)$ [17], [23]. Whereas, for the regime where K grows linearly with N (i.e., $\beta_0 > 0$), the performance of the proposed algorithms is order-wise equivalent to defective set recovery algorithms.

We also compare the uniform recovery results. The sufficient number of tests for uniform recovery as given in Corollary 1 for the algorithm **RoAI** and **CoAI** is $M = O(K \log N)$, which is significantly better than the defective set recovery algorithms, where the sufficient number of tests scale as $O(K^2 \log(\frac{N}{K}))$ [23].

- (b) Variation of M with L : Let ζ and $g(\zeta)$ be as defined above. We note that the parameter L impacts M *only* via the function $g(\zeta)$. Lemma 2 in Appendix E shows that for small values (or even moderately high values) of ζ , $g(\zeta)$ is upper bounded by an affine function in ζ . This, in turn, shows that the sufficient number of tests is also approximately affine in L ; this is also confirmed via simulation results in Section VI.
- (c) Comparison with the information theoretic lower bounds: We compare with the lower bounds on the number of tests for non-defective subset recovery, as tabulated in Table I. For the noiseless case, i.e., $u = 0, q = 0$, the sufficient number of tests are within $O(\log^2 K)$ factor of the lower bound. For the additive noise only case, the proposed algorithms incur a factor of $1/(1-q)$ increase in M . In contrast, the lower bounds indicate that the number of tests is insensitive to additive noise, when q is close to 0 (in particular, when $q < 1/K$). For the dilution noise case, the algorithms incur a factor

$\frac{1}{(1-u)}$ increase in M , which is the same as in the lower bound. We have also compared the number of tests obtained via simulations with an exact computation of the lower bounds, and, interestingly, the algorithms fall within $O(\log K)$ factor of the lower bounds; we refer the reader to Figure 4, Section VI.

- (d) Defective set recovery via non-defective subset recovery: It is interesting to note that by substituting $L = N - K$ in (3), we get $M = O\left(\frac{K \log(N-K)}{(1-u)(1-q)}\right)$, which is order-wise the same as the number of tests required for defective set identification derived in the existing literature [17], [23], [28].
- (e) Robustness under uncertainty in the knowledge of K : The theoretical guarantees presented in the above theorems hold provided the design parameter p is chosen as $O(\frac{1}{(1-u)K})$. This requires the knowledge of u and K . Note that the implementation of the recovery algorithms do not require us to know the values of K or u . These system model parameters are only required to choose the value of p for constructing the test matrix. If u and K are unknown, similar guarantees can be derived, with a penalty on the number of tests. For example, choosing p as $O(1/K)$, i.e., independent of u , results in a $\frac{1}{1-u}$ times increase in the number of tests. The impact of using an imperfect value of K can also be quantified. Let \hat{K} be the value used to design the test matrix and let $\Delta_k > 0$ be such that $\hat{K} = \Delta_k K$. That is, Δ_k parametrizes the estimation error in K . Using the fact that for large n , $(1 - \alpha/n)^n \approx \exp(-\alpha)$, it follows that with $p = O(\frac{1}{\Delta_k K})$, the number of tests increases approximately by a factor of $f_M(\Delta_k) \triangleq \Delta_k \exp\left(-(1-u)\left(\frac{1}{\Delta_k} - 1\right)\right)$ compared to the case with perfect knowledge of K , i.e., with $p = O(1/K)$. It follows that the proposed algorithms are robust to the uncertainty in the knowledge of K . For example, with $u = 0$, $f_M(1.5) = 1.09$, i.e., a 50% error in the estimation of K leads to only a 9% increase in the number of tests. Furthermore, the asymmetric nature of $f_M(\Delta_k)$ (e.g., $f_M(1.5) = 1.09$ and $f_M(0.5) = 1.3$) suggests that the algorithms are more robust when $\Delta_k > 1$ as compared to the case when $\Delta_k < 1$. We corroborate this behavior via numerical simulations also (see Table II).
- (f) Operational complexity: The execution of **RoAI** and **CoAI** requires $O(MN)$ operations, where M is the number of tests. The complexity of the LP based algorithms **RoLpAI**, **RoLpAI++** and **CoLpAI** are implementation dependent, but are, in general, much higher than **RoAI** and **CoAI**. For example, an interior-point method based implementation will require $O(N^2(M+N)^{3/2})$ operations [29]. Although this is higher than that of **RoAI** and **CoAI**, it is still attractive in comparison to the brute force search

TABLE I

FINDING A SUBSET OF L NON-DEFECTIVE ITEMS: ORDER RESULTS FOR NECESSARY NUMBER OF GROUP TESTS WHICH HOLD ASYMPTOTICALLY AS $N \rightarrow \infty$, $\frac{K}{N} \rightarrow \beta_0$, $\frac{L}{N} \rightarrow \alpha_0$ AND $\alpha_0 + \beta_0 < 1$ (SEE THEOREM 3, [21]).

No Noise ($u = 0, q = 0$)	$\Omega\left(\frac{K}{\log K} \log \frac{1-\beta_0}{1-\alpha_0-\beta_0}\right)$
Dilution Noise ($u > 0, q = 0$)	$\Omega\left(\frac{K}{(1-u) \log K} \log \frac{1-\beta_0}{1-\alpha_0-\beta_0}\right)$
Additive Noise ($u = 0, q > 0$)	$\Omega\left(\frac{K}{\min\{\log \frac{1}{q}, \log K\}} \log \frac{1-\beta_0}{1-\alpha_0-\beta_0}\right)$

based maximum likelihood methods, due to its polynomial-time complexity.

V. PROOFS OF THE MAIN RESULTS

We begin by defining some quantities and terminology that is common to all the proofs. In the following, we denote the defective set by S_d , such that $S_d \subset [N]$ and $|S_d| = K$. We denote the set of L non-defective items output by the decoding algorithm by \hat{S}_L . For a given defective set S_d , $\mathcal{E} \triangleq \{\hat{S}_L \cap S_d \neq \{\emptyset\}\}$ denotes the error event, i.e., the event that a given decoding algorithm outputs an incorrect non-defective subset and let $\Pr(\mathcal{E})$ denote its probability. Define $N_0 \triangleq (N - K) - (L - 1)$. We further let $S_z \subset [N] \setminus S_d$ denote any set of non-defective items such that $|S_z| = N_0$. Also, we let \mathcal{S}_z denote all such sets possible. Note that $|\mathcal{S}_z| = \binom{N-K}{L-1}$. Finally, recall from Lemma 1 (Section II), $\Gamma \triangleq (1 - q)(1 - (1 - u)p)^K$ and $\gamma_0 \triangleq \frac{u}{(1-(1-u)p)}$.

A. Proof of Theorem 1 and Corollary 1

The proof involves upper bounding the probability of non-defective subset recovery error of the decoding algorithms, **RoAI** and **CoAI**, and identifying the parameter regimes where they can be made sufficiently small.

For **CoAI**, recall that we compute the metric $\mathcal{T}(i) \triangleq \underline{x}_i^T \underline{y}^c - (\psi_{cb}) \underline{x}_i^T \underline{y}$ for each item i and output the set of items with the L largest metrics as the non-defective set. Clearly, for any item $i \in S_d$, if $i \in \hat{S}_L$, then there exists a set S_z of non-defective items such that for all items $j \in S_z$, $\mathcal{T}(j) \leq \mathcal{T}(i)$. Thus, for **CoAI**, it follows that,

$$\mathcal{E} \subset \bigcup_{i \in S_d} \{i \in \hat{S}_L\} \subset \bigcup_{i \in S_d} \bigcup_{S_z \in \mathcal{S}_z} \left[\bigcap_{j \in S_z} \{\mathcal{T}(j) \leq \mathcal{T}(i)\} \right]. \quad (10)$$

The algorithm **RoAI** succeeds when there exists a set of at least L non-defective items that have been tested more number of times than any of the defective items, in the tests with negative outcomes. The number of times an item i is tested in tests with negative outcomes is given by $\underline{z}(i) (= \underline{x}_i^T \underline{y}^c)$, which is

computed by **RoAI**. Hence, for any item $i \in S_d$, if $i \in \hat{S}_L$, then there exists a set S_z of non-defective items such that for all items $j \in S_z$, $\underline{z}(j) \leq \underline{z}(i)$. And, thus, (10) applies for **RoAI** also, except with \mathcal{T} replaced with \underline{z} . Also, note that $\underline{z}(i) = \mathcal{T}(i)|_{\psi_{cb}=0}$. This allows us to unify the subsequent steps in the proof for the two algorithms. We first work with the quantity $\mathcal{T}(i)$ and later specialize the results for each algorithm. The overall intuition for the proof is as follows: For any i , since $\mathcal{T}(i)$ is a sum of independent random variables, it will tend to concentrate around its mean value. For any $i \in S_d$ and $j \notin S_d$, we will show that the mean value of $\mathcal{T}(j)$ is larger than that of $\mathcal{T}(i)$. Thus, we expect the probability of the error event defined in (10) to be small.

For any $i \in S_d$ and for any $j \notin S_d$, define $\mu_i \triangleq \mathbb{E}(\mathcal{T}(i))$, $\mu_j \triangleq \mathbb{E}(\mathcal{T}(j))$, $\sigma_i^2 \triangleq \text{Var}(\mathcal{T}(i))$ and $\sigma_j^2 \triangleq \text{Var}(\mathcal{T}(j))$. It follows that,

$$\mu_j = Mp(\Gamma - \psi_{cb}(1 - \Gamma)) \quad \text{and} \quad \mu_i = Mp(\gamma_0\Gamma - \psi_{cb}(1 - \gamma_0\Gamma)) \quad (11)$$

$$\sigma_j^2 \leq Mp(\Gamma + \psi_{cb}^2(1 - \Gamma)) \quad \text{and} \quad \sigma_i^2 \leq Mp(\gamma_0\Gamma + \psi_{cb}^2(1 - \gamma_0\Gamma)). \quad (12)$$

An brief explanation of the above equations is presented in Appendix B. We note that, $(\mu_j - \mu_i) = Mp\Gamma(1 - \gamma_0)(1 + \psi_{cb}) > 0$. To simplify (10) further, we present the following proposition:

Proposition 1. Define $\tau \triangleq \frac{(\mu_j + \mu_i)}{2}$. Then, for any $\epsilon_0 > 0$ it follows that

$$Pr(\mathcal{E}) \leq K \binom{N-K}{L-1} (P_{eh})^{N_0} + KP_{ed}, \quad (13)$$

where, $P_{eh} \triangleq \mathbb{P}(\{\mathcal{T}(j) < \tau + \epsilon_0\})$ for any $j \in S_z$ and $P_{ed} \triangleq \mathbb{P}(\{\mathcal{T}(i) > \tau\})$ for any $i \in S_d$.

The proof of the above proposition is presented in Section V-A3. Note that the above definitions of P_{eh} and P_{ed} are unambiguous because the corresponding probabilities are independent of the specific choice of indices j and i , respectively.

Our next task is to bound P_{eh} and P_{ed} as defined in the above proposition. For any k , since $\mathcal{T}(k)$ is a sum of M independent random variables, each bounded by $\max(1, \psi_{cb})$, we can use Bernstein's inequality [30]³ to bound the probability of their deviation from their mean values. Since ψ_{cb} is a free parameter, we proceed by assuming that $\psi_{cb} < 1$. Thus, for any $i \in S_d$, with $\delta_0 \triangleq \tau - \mu_i = \frac{\mu_j - \mu_i}{2}$,

$$P_{ed} = P(\mathcal{T}(i) > \tau) = P(\mathcal{T}(i) > \mu_i + \delta_0) \leq \exp\left(-\frac{\delta_0^2}{2\sigma_i^2 + \frac{2}{3}\delta_0}\right). \quad (14)$$

³For ease of reference, we have stated it in Appendix G.

Similarly, for any $j \in S_z$, we choose $\epsilon_0 = \frac{\mu_j - \tau}{2} = \frac{\mu_j - \mu_i}{4}$, and get

$$P_{eh} = P(\mathcal{T}(j) < \tau + \epsilon_0) = P(\mathcal{T}(j) < \mu_j - \epsilon_0) \leq \exp\left(-\frac{\epsilon_0^2}{2\sigma_j^2 + \frac{2}{3}\epsilon_0}\right). \quad (15)$$

We now proceed separately for each algorithm to arrive at the final results. Before that, we note that by choosing $p = \frac{\alpha}{K}$ with $\alpha = \frac{1}{(1-u)}$, $\left[1 - \frac{(1-u)\alpha}{K}\right]^K \geq \exp(-2\alpha(1-u)) = e^{-2}$. This follows from the fact that for $0 < b < 1$, $(1-b) \leq e^{-b} \leq 1 - \frac{b}{2}$. Thus, $(1-q)e^{-1} \geq \Gamma \geq (1-q)e^{-2}$. We also note that $\gamma_0 < 1$ for any $u < 0.5$ and for all $K > 1$.

1) Proof for RoAI

For **RoAI**, $\psi_{cb} = 0$. Thus, from (11) and (12) we have, $\mu_j - \mu_i = Mp\Gamma(1 - \gamma_0)$, $\sigma_j^2 \leq Mp\Gamma$ and $\sigma_i^2 \leq Mp\gamma_0\Gamma$. Recall, $\delta_0 = \frac{\mu_j - \mu_i}{2}$ and $\epsilon_0 = \frac{\mu_j - \mu_i}{4}$. Note that, $2\sigma_i^2 + (2/3)\delta_0 < Mp\Gamma(2\gamma_0 + (1 - \gamma_0)/3) < 2Mp\Gamma$. Similarly, $2\sigma_j^2 + (2/3)\epsilon_0 < Mp\Gamma(2 + (1 - \gamma_0)/6) < 3Mp\Gamma$. Thus, from (14) and (15), we have

$$P_{ed} \leq \exp\left(-\frac{Mp\Gamma(1 - \gamma_0)^2}{8}\right) \text{ and } P_{eh} \leq \exp\left(-\frac{Mp\Gamma(1 - \gamma_0)^2}{48}\right). \quad (16)$$

Thus, choosing $p = \frac{1}{(1-u)K}$ and noting that $\Gamma \geq e^{-2}(1-q)$, from (13) we get,

$$\mathbb{P}(\mathcal{E}) \leq \exp\left[-\frac{M(1 - \gamma_0)^2(1 - q)N_0}{C_{a1}K(1 - u)} + \log\left(K\binom{N - K}{L - 1}\right)\right] + \exp\left[-\frac{M(1 - \gamma_0)^2(1 - q)}{C_{a2}K(1 - u)} + \log K\right],$$

with $C_{a1} = 48e^2$ and $C_{a2} = 8e^2$. Thus, if M is chosen as specified in (3), with the constants C_{a1} , C_{a2} chosen as above, then the error probability is upper bounded by $\exp(-c_0 \log[K\binom{N-K}{L-1}]) + \exp(-c_0 \log K)$.

2) Proof for CoAI

We first bound P_{ed} . With $\psi_{cb} = \psi_0$, where $\psi_0 \triangleq \frac{\gamma_0\Gamma}{1 - \gamma_0\Gamma}$, we have $\sigma_i^2 \leq Mp\gamma_0\Gamma(1 + \psi_0)$. Also, we note that $\psi_0 < 1$. Thus, $2\sigma_i^2 + (2/3)\delta_0 < Mp\Gamma(1 + \psi_0)(2\gamma_0 + (1 - \gamma_0)/3) < 2Mp\Gamma(1 + \psi_0)$. Thus, from (14), we get

$$P_{ed} \leq \exp\left(-\frac{Mp\Gamma(1 + \psi_0)(1 - \gamma_0)^2}{8}\right). \quad (17)$$

With ψ_0 as above, it follows that, $2\sigma_j^2 + (2/3)\epsilon_0 < Mp\Gamma\left(2 + 2\frac{\gamma_0^2\Gamma}{(1 - \gamma_0\Gamma)} + \frac{(1 - \gamma_0)(1 + \psi_0)}{6}\right) < 3Mp\Gamma(1 + \psi_0)$, since $1 + \psi_0 = \frac{1}{1 - \gamma_0\Gamma}$. Thus, from (15), we get

$$P_{eh} \leq \exp\left(-\frac{Mp\Gamma(1 + \psi_0)(1 - \gamma_0)^2}{48}\right). \quad (18)$$

The next steps follow exactly as for **RoAI** and if M is chosen as specified in (3), with the constant C_{a1} and C_{a2} chosen as $48e^2$ and $8e^2$, respectively, then the error probability remains smaller than $\exp(-c_0 \log[K\binom{N-K}{L-1}]) + \exp(-c_0 \log K)$.

3) Proof of Proposition 1

For $i \in S_d$, define $\mathcal{H}_i \triangleq \{\mathcal{T}(i) \leq \tau\}$. The error event in (10) is a subset of the right hand side in the following equation:

$$\mathcal{E} \subset \bigcup_{i \in S_d} \left(\left\{ \bigcup_{S_z \in \mathcal{S}_z} \bigcap_{j \in S_z} (\mathcal{E}_{ij} \cap \mathcal{H}_i) \right\} \cup \overline{\mathcal{H}_i} \right), \quad (19)$$

where $\mathcal{E}_{ij} \triangleq \{\mathcal{T}(j) \leq \mathcal{T}(i)\}$ for any $i \in S_d$ and $j \in S_z$. In the above, we have used the fact that, for any two sets A and B , $A \subset \{A \cap B\} \cup \overline{B}$. Further, using monotonicity properties, we have

$$\{\mathcal{E}_{ij} \cap \mathcal{H}_i\} \subset \{\mathcal{T}(j) \leq \tau\} \subset \{\mathcal{T}(j) < \tau + \epsilon_0\}, \quad (20)$$

where $\epsilon_0 > 0$ is any constant. Consider any non-defective item $j \in S_z$. We note that for any given \underline{y} , $\mathcal{T}(j)$ can be represented as a function of only \underline{x}_j , i.e., the j^{th} column of the test matrix \mathbf{X} . From (1), since $j \notin S_d$, the output is independent of the entries of \underline{x}_j . Hence, for all $j \notin S_d$, and hence for all $j \in S_z$, $\mathcal{T}(j)$'s are independent. Using this observation, the claim in the proposition now follows from (19) and (20) by accounting for the cardinalities of different sets involved in the union bounding.

4) Proof of Corollary 1

For the uniform case, we use the union bound over all possible choices of the defective set. The proof of the corollary follows same steps as the proof of Theorem 1; the only difference comes on account of the additional union bounding that has to be done to account for all possible choices of the defective set. Here, we briefly discuss the different multiplicative factors that have to be included because of this additional union bound. Let \mathcal{S}_d denote the set of all possible defective sets. Note that $|\mathcal{S}_d| = \binom{N}{K}$. From (19) in the proof of Proposition 1, we note that

$$\mathcal{E} \subset \left\{ \bigcup_{S_d \in \mathcal{S}_d} \bigcup_{i \in S_d} \bigcup_{S_z \in \mathcal{S}_z} \bigcap_{j \in S_z} (\mathcal{E}_{ij} \cap \mathcal{H}_i) \right\} \cup \left\{ \bigcup_{S_d \in \mathcal{S}_d} \bigcup_{i \in S_d} \overline{\mathcal{H}_i} \right\}, \quad (21)$$

Thus, for the first term in (13), an additional multiplicative factor of $\binom{N}{K}$ is needed to account for all possible defective sets. For the second term, we note that

$$\bigcup_{S_d \in \mathcal{S}_d} \bigcup_{i \in S_d} \overline{\mathcal{H}_i} \subseteq \bigcup_{i \in [N]} \overline{\mathcal{H}_i}. \quad (22)$$

Thus, for the second term in (13), the multiplicative factor of K in (13) gets replaced by a factor of N , and no additional combinatorial multiplicative factors are needed. The corollary now follows using the same steps as in the proof of **RoAI** and **CoAI**.

B. Proof of Theorem 2

Let $\mathbf{X} \in \{0, 1\}^{M \times N}$ denote the random test matrix, $\underline{y} \in \{0, 1\}^M$ the output of the group test, $Y_z \triangleq \{l \in [M] : \underline{y}(l) = 0\}$ with $M_z \triangleq |Y_z|$, and $Y_p \triangleq \{l \in [M] : \underline{y}(l) = 1\}$ with $M_p \triangleq |Y_p|$. Let $\mathbf{X}_z \triangleq \mathbf{X}(Y_z, :)$ and $\mathbf{X}_p \triangleq \mathbf{X}(Y_p, :)$. Note that $\mathbf{X}_z \in \{0, 1\}^{M_z \times N}$ and $\mathbf{X}_p \in \{0, 1\}^{M_p \times N}$. For the ease of performance analysis of the LP described in (5), we work with the following equivalent program:

$$\begin{aligned}
 & \underset{\underline{z}}{\text{minimize}} && \underline{1}_{M_z}^T \mathbf{X}_z \underline{z} \\
 \text{(LP0a)} \quad & \text{subject to} && \underline{0}_N \preceq \underline{z} \preceq \underline{1}_N, \\
 & && \underline{1}_N^T \underline{z} \geq (N - L).
 \end{aligned} \tag{23}$$

The above formulation has been arrived at by eliminating the equality constraints and replacing the optimization variable \underline{z} by $(\underline{1}_N - \underline{z})$. Hence, the non-defective subset output by (23) is indexed by the *smallest* L entries in the solution of (LP0a) (as opposed to largest L entries in the solution of (LP0)). We know that strong duality holds for a linear program and that any pair of primal and dual optimal points satisfy the Karush-Kuhn-Tucker (KKT) conditions [31]. Hence, a characterization of the primal solution can be obtained in terms of the dual optimal points by using the KKT conditions. Let $\underline{\lambda}_1, \underline{\lambda}_2 \in \mathbb{R}^N$ and $\nu \in \mathbb{R}$ denote the dual variables associated with the inequality constraints in (LP0a). The KKT conditions for any pair of primal and dual optimal points corresponding to (LP0a) can be written as follows:

$$\underline{1}_{M_z}^T \mathbf{X}_z - \underline{\lambda}_1 + \underline{\lambda}_2 - \nu \underline{1}_N = \underline{0}_N \tag{24}$$

$$\underline{\lambda}_1 \circ \underline{z} = \underline{0}_N; \quad \underline{\lambda}_2 \circ (\underline{z} - \underline{1}_N) = \underline{0}_N; \quad \nu(\underline{1}_N^T \underline{z} - (N - L)) = 0; \tag{25}$$

$$\underline{0}_N \preceq \underline{z} \preceq \underline{1}_N; \quad \underline{1}_N^T \underline{z} \geq (N - L); \quad \underline{\lambda}_1 \succeq \underline{0}_N; \quad \underline{\lambda}_2 \succeq \underline{0}_N; \quad \nu \geq 0; \tag{26}$$

Let $(\underline{z}, \underline{\lambda}_1, \underline{\lambda}_2, \nu)$ be the primal, dual optimal point, i.e., a point satisfying the set of equations (24)-(26). Let S_d denote the set of defective items. Further, let \hat{S}_L denote the index set corresponding to the smallest L entries, and hence the declared set of non-defective items, in the primal solution \underline{z} . We first derive a sufficient condition for successful non-defective subset recovery with **RoLpAl**.

Proposition 2. *If $\underline{\lambda}_2(i) > 0 \forall i \in S_d$, then $\hat{S}_L \cap S_d = \{\emptyset\}$.*

Proof: See Appendix C.

Let \mathcal{E} , $\mathbb{P}(\mathcal{E})$, S_z and \mathcal{S}_z be as defined at the beginning of this section. The above sufficiency condition for successful non-defective subset recovery, in turn, leads to the following:

Proposition 3. *The error event associated with **RoLpAI** satisfies:*

$$\mathcal{E} \subseteq \bigcup_{i \in S_d} \bigcup_{S_z \in \mathcal{S}_z} \{ \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i) \geq \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, j), \forall j \in S_z \}. \quad (27)$$

Proof: Define $\mathcal{E}_0(i) \triangleq \{\lambda_2(i) = 0\}$. We first note, from (24), that for any $i \in [N]$

$$\lambda_2(i) = 0 \implies \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i) = \lambda_1(i) + \nu \geq \nu. \quad (28)$$

Define $\theta_0 \triangleq \max_{\{i: \lambda_1(i)=0\}} \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i)$ and $\theta_1 \triangleq \min_{\{i: \lambda_1(i)>0\}} \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i)$. We relate θ_0 , θ_1 and ν as follows:

Proposition 4. *The dual optimal variable ν satisfies $\theta_0 \leq \nu < \theta_1$.*

Proof: See Appendix D.

From the above proposition and (28) it follows that

$$\mathcal{E}_0(i) \subseteq \{ \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i) \geq \theta_0 \}. \quad (29)$$

We note that there exists at most L items for which $\lambda_1(i) > 0$; otherwise the solution would violate the primal feasibility constraint: $\underline{\mathbf{1}}_N^T \mathbf{z}(i) \geq (N - L)$. Thus, there exist at least $(N - K) - (L - 1)$ non-defective items in the set $\{i : \lambda_1(i) = 0\}$. From (29), there exists a set S_z of $(N - K) - (L - 1)$ non-defective items such that $\{ \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i) \geq \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, j), \forall j \in S_z \}$. Taking the union bound over all possible S_z , we get

$$\mathcal{E}_0(i) \subseteq \bigcup_{S_z \in \mathcal{S}_z} \{ \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i) \geq \underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, j), \forall j \in S_z \}, \quad (30)$$

and (27) now follows since using Proposition 2 we have, $\mathcal{E} \subseteq \bigcup_{i \in S_d} \mathcal{E}_0(i)$. \blacksquare

Note that, for a given i , the quantity $\underline{\mathbf{1}}_{M_z}^T \mathbf{X}_z(:, i)$ is the same as the quantity $\mathcal{T}(i)$ with $\psi_{cb} = 0$ as defined in the proof of Theorem 1, and (27) is the same as (10). Thus, following the same analysis as in Section V-A, it follows that, if M satisfies (3) with $\psi_0 = 0$, the LP relaxation based algorithm **RoLpAI** succeeds in recovering L non-defective items with probability exceeding $1 - \exp\left(-c_0 \log \left[K \binom{N-K}{L-1} \right]\right) + \exp(-c_0 \log K)$.

C. Proof Sketch for Theorem 3

We use the same notation as in Theorem 2 and analyze an equivalent program that is obtained by replacing $(1 - \underline{z})$ by \underline{z} . We note that **LP2** differs from **LP0** only in terms of the objective function, and the constraint set remains the same. And thus, the complimentary slackness and the primal dual feasibility conditions are the same as given in (25) and (26), respectively. The zero gradient condition for **LP2** is

given by:

$$\underline{1}_{M_z}^T \mathbf{X}_z - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p - \underline{\lambda}_1 + \underline{\lambda}_2 - \nu \underline{1}_N = \underline{0}_N. \quad (31)$$

Let the error event associated with **CoLpAI** be denoted by \mathcal{E} . Let $i \in S_d$, and define $\mathcal{E}_i \triangleq \{i \in \hat{S}_L\}$. Note that $\mathcal{E} \subseteq \bigcup_{i \in S_d} \mathcal{E}_i$. Further, it follows that $\mathcal{E}_i \subseteq \mathcal{A}_i \cup \mathcal{B}_i$, where $\mathcal{A}_i \triangleq \{\lambda_2(i) = 0\}$ and $\mathcal{B}_i \triangleq \{\mathcal{E}_i \cap \{\lambda_2(i) > 0\}\}$. Let us first analyze \mathcal{B}_i . Using similar arguments as in Propositions 2 and 4, it can be shown that,

$$\mathcal{B}_i \subseteq \{\nu = 0\} \subseteq \bigcup_{S_z \in \mathcal{S}_z} \left\{ \underline{1}_{M_z}^T \mathbf{X}_z(:, j) - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p(:, j) \leq 0, \forall j \in S_z \right\}, \quad (32)$$

where $S_z \subset [N] \setminus S_d$ is any set of non-defective items such that $|S_z| = (N - K) - (L - 1)$ and \mathcal{S}_z denotes all such sets possible. Further, using similar arguments as in the proof of Theorem 2, it can be shown that

$$\mathcal{A}_i \subseteq \bigcup_{S_z \in \mathcal{S}_z} \left\{ \underline{1}_{M_z}^T \mathbf{X}_z(:, i) - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p(:, i) \geq \underline{1}_{M_z}^T \mathbf{X}_z(:, j) - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p(:, j), \forall j \in S_z \right\}, \quad (33)$$

where S_z and \mathcal{S}_z are as defined above.

The subsequent analysis follows by using the Bernstein inequality to upper bound the probability of events \mathcal{A}_i and \mathcal{B}_i in a manner similar to the previous proofs; we omit the details for the sake of brevity. Define $\psi'_0 \triangleq \min\left(\frac{\gamma_0 \Gamma}{1 - \gamma_0 \Gamma}, \frac{\Gamma}{2(1 - \Gamma)}\right)$. Note that, with $\psi_{lp} = \psi'_0$, $\mathbb{E}([\underline{1}_{M_z}^T \mathbf{X}_z(:, j) - \psi_{lp} \underline{1}_{M_p}^T \mathbf{X}_p(:, j)]) \geq Mp\Gamma/2 > 0$ for any $j \in S_z$. This helps in upper bounding the probability of \mathcal{B}_i using Bernstein's inequality. In essence, it can be shown that there exists an absolute constant $C_{4b} > 0$, such that

$$\mathbb{P}\left(\bigcup_{i \in S_d} \mathcal{B}_i\right) \leq \exp\left(-\left\{\frac{Mp\Gamma N_0}{C_{4b}} - \log\left[K\binom{N-K}{L-1}\right]\right\}\right). \quad (34)$$

Similarly, following the same steps as in the proof of Theorem 1, it can be shown that, for the chosen value of ψ_{lp} , there exists an absolute constant C_{4a} such that

$$\mathbb{P}\left(\bigcup_{i \in S_d} \mathcal{A}_i\right) \leq \exp\left(-\frac{Mp\Gamma(1 - \gamma_0)^2 N_0}{C_{4a}} + \log\left[K\binom{N-K}{L-1}\right]\right) + \exp\left(-\frac{Mp\Gamma(1 - \gamma_0)^2}{C_{4c}} + \log K\right). \quad (35)$$

The final result now follows by substituting $p = \frac{1}{(1-u)K}$, since, by choosing M as in (3) with $\psi_0 = 0$, $C_{a1} = \max\{C_{4a}, C_{4b}\}$ and $C_{a2} = C_{4c}$, the total error in (34), (35) can be upper bounded as $2 \exp\left(-c_0 \log\left(K\binom{N-K}{L-1}\right)\right) + \exp(-c_0 \log K)$. This concludes the proof.

VI. SIMULATIONS

In this section, we investigate the empirical performance of the algorithms proposed in this work for non-defective subset recovery. In contrast to the previous section, where theoretical guarantees on

the number of tests were derived based on the analysis of the upper bounds on probability of error of these algorithms, here we find the exact number of tests required to achieve a given performance level, thus highlighting the practical ability of the proposed algorithms to recover a non-defective subset. This, apart from validating the general theoretical trends, also facilitates a direct comparison of the presented algorithms.

Our setup is as follows. For a given set of operating parameters, i.e., N , K , u , q and M , we choose a defective set $S_d \subset [N]$ randomly such that $|S_d| = K$ and generate the test output vector \underline{y} according to (1). We then recover a subset of L non-defective items using the different recovery algorithms, i.e., **RoAl**, **CoAl**, **RoLpAl**, **RoLpAl++** and **CoLpAl**, and compare it with the defective set. The empirical probability of error is set equal to the fraction of the trials for which the recovery was not successful, i.e., the output non-defective subset contained at least one defective item. This experiment is repeated for different values of M and L . For each trial, the test matrix \mathbf{X} is generated with random Bernoulli i.i.d. entries, i.e., $\mathbf{X}_{ij} \sim \mathcal{B}(p)$, where $p = 1/K$. Also, for **CoAl** and **CoLpAl**, we set $\psi_{cb} = \frac{\gamma_0 \Gamma}{1 - \gamma_0 \Gamma}$ and $\psi_{lp} = \min\left(\frac{\gamma_0 \Gamma}{1 - \gamma_0 \Gamma}, \frac{\Gamma}{2(1 - \Gamma)}\right)$, respectively. Unless otherwise stated, we set $N = 256$, $K = 16$, $u = 0.05$, $q = 0.1$ and we vary L and M .

Figure 2 shows the variation of the empirical probability of error with the number of tests, for $L = 64$ and $L = 128$. These curves demonstrate the theoretically expected exponential behavior of the average error rates, the similarity of the error rate performance of algorithms **RoAl** and **RoLpAl**, and the performance improvement offered by **RoLpAl++** at higher values of L . We also note that, as expected, the algorithms that use tests with both positive and negative outcomes perform better than the algorithms that use only tests with negative outcomes.

Figure 3 presents the number of tests M required to achieve a target error rate of 10% as a function of the size of the non-defective subset, L . We note that for small values of L , the algorithms perform similarly, but, in general, **CoAl** and **CoLpAl** are the best performing algorithms across all values of L . We also note that, as argued in Section III-C, **RoLpAl++** performs similar to **RoLpAl** for small values of L and for large values of L the performance of the former is the same as that of **CoLpAl**. Also, as mentioned in Section IV, we note the linear increase in M with L , especially for small values of L . We also compare the algorithms proposed in this work with an algorithm that identifies the non-defective items by first identifying the defective items, i.e., we compare the “direct” and “indirect” approach [21] of

identifying a non-defective subset. We first employ a defective set recovery algorithm for identifying the defective set and then choose L items uniformly at random from the complement set. This algorithm is referred to as “InDirAl” in Figure 3. In particular, we have used “**No-LiPo-**” algorithm [17] for defective set identification. It can be easily seen that the “direct” approach significantly outperforms the “indirect” approach. We also compare against a non-adaptive scheme that tests items one-by-one. The item to be tested in each test is chosen uniformly at random from the population. We choose the top L items tested in all the tests with negative outcomes as the non-defective subset. This algorithm is referred to as “NA1by1” (Non-Adaptive 1-by-1) in Figure 3. It is easy to see that the group testing based algorithms significantly outperform the NA1by1 strategy.

Figure 4 compares the number of tests required to achieve a target error rate of 10% for **CoAl** with the information theoretic lower bound for two different values of K .⁴ It can be seen that the empirical performance of **CoAl** is within $O(\log K)$ of the lower bound. The performance of the other algorithms is found to obey a similar behavior.

As discussed in Section IV, the parameter settings require the knowledge of K . Here, we investigate the sensitivity of the algorithms on the test matrix designed assuming a nominal value of K to mismatches in its value. Let the true number of defective items be K_t . Let $M(\hat{K}, K_t)$ denote the number of tests required to achieve a given error rate when the test is designed with $K = \hat{K}$. Let $\Delta_M(\hat{K}, K_t) \triangleq \frac{M(\hat{K}, K_t)}{M(K_t, K_t)}$. Thus, $\Delta_M(\hat{K}, K_t)$ represents the penalty paid compared to the case when the test is designed knowing the number of defective items. Table II shows the empirically computed Δ_M for different values of uncertainty factor $\Delta_K \triangleq \frac{\hat{K}}{K_t}$ for the different algorithms. We see that the algorithms exhibit robustness to the uncertainty in the knowledge of K . For example, even when $\hat{K} = 2K_t$, i.e., $\Delta_K = 2$, we only pay a penalty of approximately 17% for most of the algorithms. Also, as suggested by the analysis of the upper bounds in Section IV, the algorithms exhibit asymmetric behavior in terms of robustness and are more robust for $\Delta_K > 1$ compared to when $\Delta_K < 1$.

Figure 5 shows the performance of different algorithms with the variations in the system noise parameters. Again, in agreement with the analysis of the probability of error, the algorithms perform

⁴We refer the reader to Theorem 3 and Section IV in [21] for a detailed discussion on the information theoretic lower bound. Also, see equations (7) and (9) in [32] for the derivation of the mutual information term that is required for computing the lower bound for the group testing signal model.

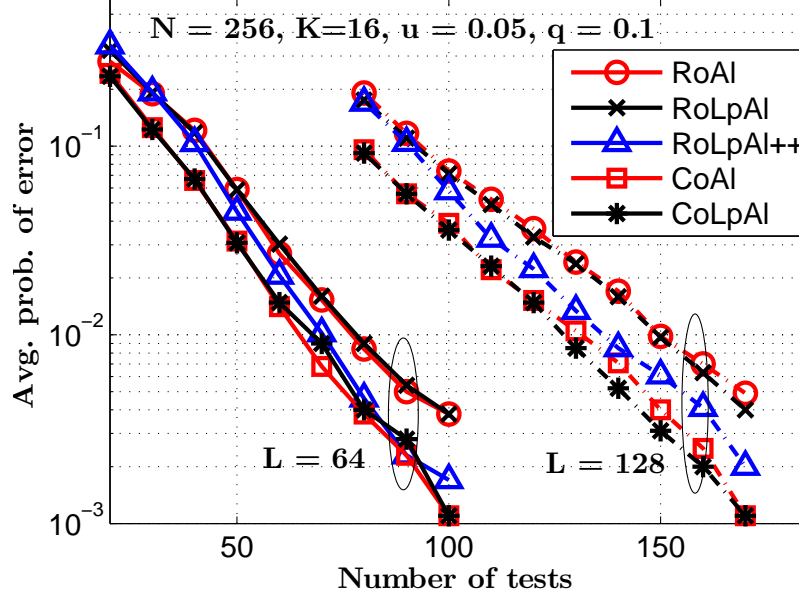


Fig. 2. Average probability of error (APER) vs. number of tests M . The APER decays exponentially with M .

TABLE II

ROBUSTNESS OF THE NON-DEFECTIVE SUBSET IDENTIFICATION ALGORITHMS TO UNCERTAINTY IN THE KNOWLEDGE OF K . THE NUMBERS IN THE TABLE ARE $\Delta_M(\hat{K}, K_t)$.

$K_t = 16, N = 256, L = 128, q = 0.1, u = 0.05$			
	$\Delta_K = 0.75$	$\Delta_K = 1.5$	$\Delta_K = 2.0$
RoAI	1.13	1.06	1.20
CoAI	1.13	1.04	1.17
RoLpAI	1.09	1.04	1.17
RoLpAI++	1.04	1.00	1.17
CoLpAI	1.11	1.03	1.19

similarly with respect to variations in both the additive and dilution noise.

VII. CONCLUSIONS

In this work, we have proposed analytically tractable and computationally efficient algorithms for identifying a non-defective subset of a given size in a noisy non-adaptive group testing setup. We have derived upper bounds on the number of tests for guaranteed correct subset identification and we have shown that the upper bounds and information theoretic lower bounds are order-wise tight up to a poly-

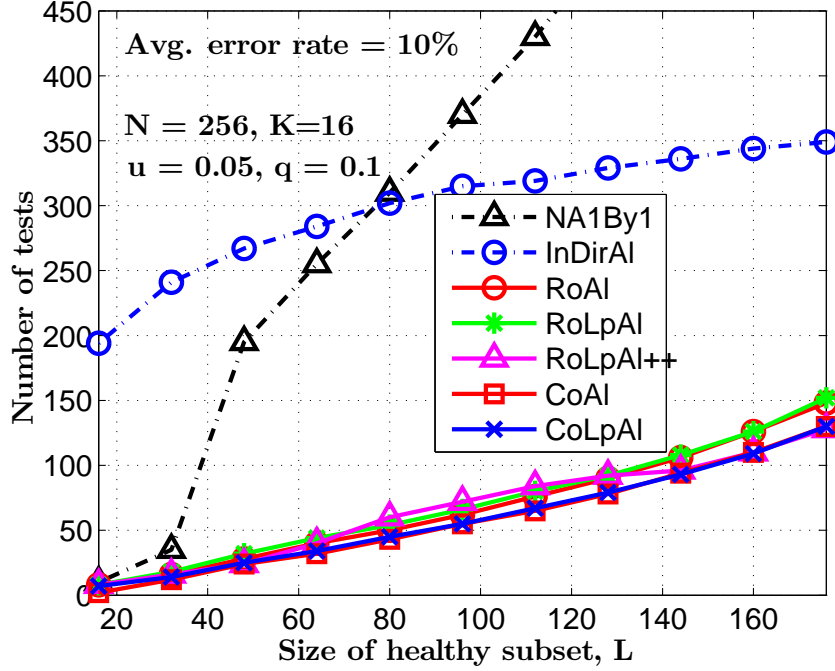


Fig. 3. Number of tests vs. size of non-defective subset. Algorithm **CoLpAI** performs the best among the ones considered. The direct approach for finding non-defective items significantly outperforms both the indirect approach (“InDirAI”), where defective items are identified first and the non-defective items are subsequently chosen from the complement set [21], as well as the item-by-item testing approach (“NA1By1”).

log factor. We have shown that the algorithms are robust to the uncertainty in the knowledge of system parameters. Also, it was found that the algorithms that use both positive and negative outcomes, namely **CoAI** and the LP relaxation based **CoLpAI**, gave the best performance for a wide range of values of L , the size of non-defective subset to be identified. In this work, we have considered the randomized pooling strategy. It will be interesting to study deterministic constructions for the purpose of non-defective subset identification; this could be considered in a future extension of this work. Another interesting question to investigate is to extend the non-defective subset identification problem to scenarios with structured pooling strategies, e.g., for graph constrained group testing where the pools are constrained by the nodes that lie on a path of a given graph.

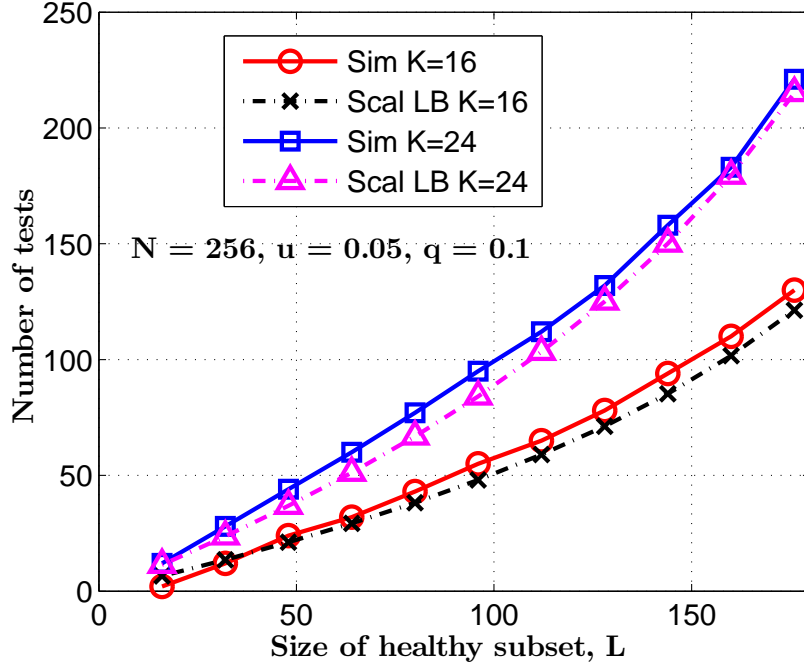


Fig. 4. Comparison of **CoAI** with the *scaled* information theoretic lower bounds. Here, the lower bounds have been scaled by a multiplicative factor of $\log(K)$. The close agreement of the scaled lower bound with the performance of the algorithm shows that **CoAI** is within a $\log(K)$ factor of the lower bounds.

APPENDIX

A. Proof of Lemma 1

We note that a test outcome is 0 only if none of the K defective items participate in the test and the output is not corrupted by the additive noise. (a) now follows by noting that the probability that an item does not participate in the group test is given by $(1-p) + pu$. (b) follows from (1). For (c) we note that, given that $X_{li} = 1$ for any $i \in S_d$, the outcome is 0 only if the i^{th} item does not participate in the test (despite $X_{li} = 1$) and none of the remaining $K-1$ defective items participate in the test (either the entry of the test matrix is zero or the item gets diluted out by noise) and the test outcome is not corrupted by additive noise. That is, $\mathbb{P}(Y_l = 0 | X_{li} = 1) = u(1 - (1-p)u)^{K-1}(1-q) = \gamma_0 \Gamma$. The other part follows similarly. (d) follows by noting that for any $i \in S_d$ and $j \notin S_d$, $\mathbb{P}(Y_l | X_{li}, X_{lj}) = \mathbb{P}(Y_l | X_{li})$. By Bayes rule and part (b) in this lemma, we get: $\mathbb{P}(X_{li}, X_{lj} | Y_l) = \frac{\mathbb{P}(Y_l | X_{li}, X_{lj}) \mathbb{P}(X_{li}) \mathbb{P}(X_{lj})}{\mathbb{P}(Y_l)} = \mathbb{P}(X_{li} | Y_l) \mathbb{P}(X_{lj} | Y_l)$. Hence the proof.

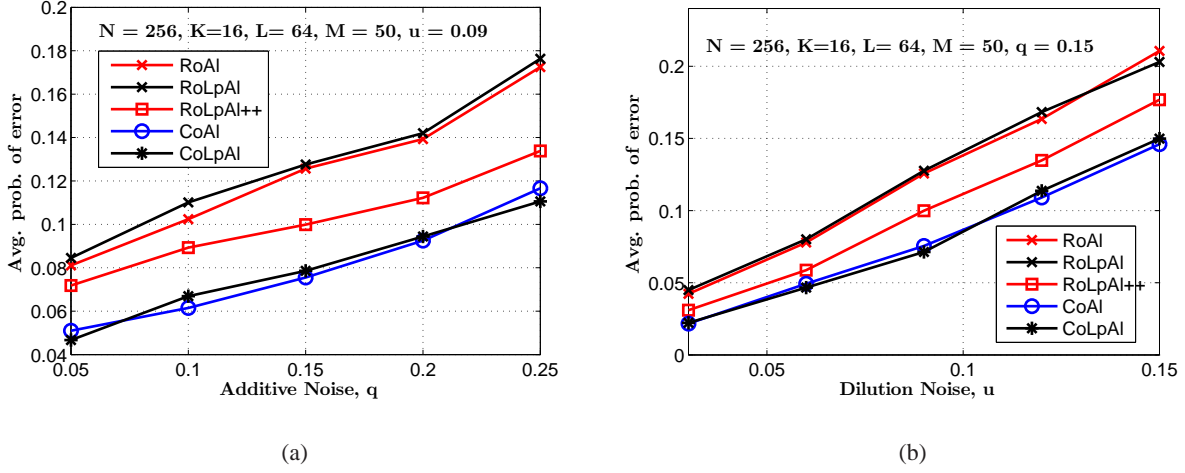


Fig. 5. Variation of the average probability with (a) additive noise (q) and (b) dilution noise (u).

B. Proof of (11) and (12)

For $i \in S_d$ and $j \notin S_d$, let $\underline{x}_j(l) \triangleq X_{jl}$, $\underline{x}_i(l) \triangleq X_{il}$ and $\underline{y}(l) \triangleq Y_l$. For any $k \in [N]$, we note that $\mathcal{T}(k)$ can be written as a sum of M independent random variables $\sum_{l=1}^M Z_{kl}$, where Z_{kl} takes value 1 with probability $\mathbb{P}(X_{kl} = 1, Y_l = 0)$, $-\psi_{cb}$ with probability $\mathbb{P}(X_{kl} = 1, Y_l = 1)$, and takes the value 0 otherwise. From Lemma 1, we know that $P(Y_l = 0 | X_{il} = 1) = \gamma_0 \Gamma$ and $P(Y_l = 0 | X_{jl} = 1) = \Gamma$ and thus (11) follows. Further, (12) follows by noting that

$$\text{Var}(Z_{jl}) \leq \mathbb{E}(Z_{jl}^2) = p(\Gamma + \psi_{cb}^2(1 - \Gamma))$$

$$\text{Var}(Z_{il}) \leq \mathbb{E}(Z_{il}^2) = p(\gamma_0 \Gamma + \psi_{cb}^2(1 - \gamma_0 \Gamma)).$$

C. Proof of Proposition 2

We first prove that, for all $i \in \hat{S}_L$, $\underline{\lambda}_2(i) = 0$. The proof is based on contradiction. Suppose $\exists j \in \hat{S}_L$ such that $\underline{\lambda}_2(j) > 0$. This implies, from the complimentary slackness conditions (25), $\underline{z}(j) = 1$ and thus, $\underline{\lambda}_1(j) = 0$. Since j^{th} item is amongst the smallest L entries, this implies that $\underline{1}_N^T \underline{z} > (N - L)$. Hence, $\nu = 0$. From the zero gradient condition in (24), it follows that $\underline{1}_{M_z}^T \mathbf{X}_z(:, j) = -\underline{\lambda}_2(j) < 0$, which is not possible, as all entries in \mathbf{X} are nonnegative. It then follows that $\forall i \in \hat{S}_L$ $\underline{\lambda}_2(i) = 0$. Thus, if $\underline{\lambda}_2(i) > 0 \forall i \in S_d$, then these items cannot belong to the first L entries in the primal solution \underline{z} , i.e., $S_d \cap \hat{S}_L = \{\emptyset\}$.

D. Proof of Proposition 4

Suppose $\nu < \theta_0$. Then $\exists i$ such that $\underline{\lambda}_1(i) = 0$ and $\nu < \underline{1}_{M_z}^T \mathbf{X}_z(:, i)$. Thus, from (24), $\underline{\lambda}_2(i) = \nu - \underline{1}_{M_z}^T \mathbf{X}_z(:, i) < 0$, which violates the dual feasibility conditions (26). Thus, $\nu \geq \theta_0$. Similarly, let $\nu \geq \theta_1$. Then $\exists i$ such that $\underline{\lambda}_1(i) = 1$ and $\nu \geq \underline{1}_{M_z}^T \mathbf{X}_z(:, i)$. Thus, from (24), $\underline{\lambda}_2(i) = \underline{\lambda}_1(i) + \nu - \underline{1}_{M_z}^T \mathbf{X}_z(:, i) \geq 1$, which is a contradiction since $\underline{\lambda}_1(i) > 0$ implies $\underline{\lambda}_2(i) = 0$. Thus, $\nu \geq \theta_1$ is not possible.

E. Affine characterization of the function $\frac{H_b(\alpha)}{1-\alpha}$

Lemma 2. Let $H_b(\cdot)$ represent the binary entropy function. Then, for $0 < \alpha \leq \alpha_h < 1$, there exist positive absolute constants $c_0, c_1 > 0$, with c_1 depending on α_h , such that

$$\frac{H_b(\alpha)}{1-\alpha} \leq c_0\alpha + c_1. \quad (36)$$

To establish (36), we note that

$$\begin{aligned} \frac{H_b(\alpha)}{1-\alpha} &= -\frac{\alpha}{1-\alpha} \log(\alpha) - \log(1-\alpha) = \frac{\alpha}{1-\alpha} \sum_{i=1}^{\infty} \frac{(1-\alpha)^i}{i} + \sum_{i=1}^{\infty} \frac{\alpha^i}{i} \\ &\leq \alpha \left(1 + \frac{(1-\alpha)}{2} + \frac{(1-\alpha)^2}{3} \right) + \frac{\alpha(1-\alpha)^3}{4} \left(\sum_{i=1}^{\infty} (1-\alpha)^{i-1} \right) \\ &\quad + \alpha + \frac{\alpha^2}{2} + \frac{\alpha^3}{3} + \frac{\alpha^4}{4} \left(\sum_{i=1}^{\infty} \alpha^{i-1} \right) \\ &\leq \frac{17}{6}\alpha + \frac{1}{4} \left[(1-\alpha)^3 + \frac{\alpha^4}{1-\alpha} \right] \leq c_0\alpha + c_1, \end{aligned}$$

where $c_0 = 17/6$ and c_1 is obtained by appropriately bounding the second term when $0 \leq \alpha \leq \alpha_h$. In particular, for $\alpha_h \leq 0.5$, $c_1 = 0.25$ will satisfy (36).

F. Discussion on the theoretical guarantees for **RoLpAl++**

The discussion for **RoLpAl++** proceeds on similar lines as **RoLpAl**. We use the same notation as in Section V-B, and, as before, we analyze an equivalent LP obtained by eliminating the equality constraints and substituting $(1 - \underline{z})$ by \underline{z} . The corresponding KKT conditions for a pair of primal and dual optimal points are as follows:

$$\underline{1}_{M_z}^T \mathbf{X}_z - \underline{\mu}^T \mathbf{X}_p - \underline{\lambda}_1 + \underline{\lambda}_2 - \nu \underline{1}_N = \underline{0}_N \quad (37)$$

$$\underline{\mu} \circ (\mathbf{X}_p \underline{z} - (1 - \epsilon_0) \underline{1}_{M_p}) = \underline{0}_{M_p}; \quad \underline{\lambda}_1 \circ \underline{z} = \underline{0}_N; \quad \underline{\lambda}_2 \circ (\underline{z} - \underline{1}_N) = \underline{0}_N; \quad \nu(\underline{1}_N^T \underline{z} - (N - L)) = 0; \quad (38)$$

$$\underline{0}_N \preceq \underline{z} \preceq \underline{1}_N; \quad \underline{1}_N^T \underline{z} \geq (N - L); \quad \underline{\mu} \succeq \underline{0}_{M_p}; \quad \underline{\lambda}_1 \succeq \underline{0}_N; \quad \underline{\lambda}_2 \succeq \underline{0}_N; \quad \nu \geq 0; \quad (39)$$

In the above, $\underline{\mu} \in \mathbb{R}^{M_p}$ is the dual variable associated with constraint (8) of **LP1**. Let $(\underline{z}, \underline{\mu}, \underline{\lambda}_1, \underline{\lambda}_2, \nu)$ be a primal, dual optimal point satisfying the above equations. We first prove the following:

Proposition 5. *If $\lambda_2(i) > 0$, then $\underline{\mu}^T \mathbf{X}_p(:, i) = 0$.*

Proof: For any $l \in [M_p]$, if $\mathbf{X}_p(l, i) = 0$ then $\underline{\mu}(l) \mathbf{X}_p(l, i) = 0$. If $\mathbf{X}_p(l, i) = 1$, then for the l^{th} test $\mathbf{X}_p(l, :)\underline{z} \geq 1 > (1 - \epsilon_0)$, since $\lambda_2(i) > 0$ implies $z(i) = 1$. This implies $\underline{\mu}(l) = 0$, and thus $\underline{\mu}(l) \mathbf{X}_p(l, i) = 0$. Hence the proposition follows. \blacksquare

Using the above, it is easy to see that Proposition 2 holds in this case also. Furthermore, using the same arguments as in Section V-B, it can be shown that the error event associated with **RoLpAl++**, \mathcal{E} , satisfies $\mathcal{E} \subseteq \cup_{i \in S_d} \cup_{S_z \in \mathcal{S}_z} \{\cap_{j \in S_z} \mathcal{E}_0(i, j)\}$, where

$$\mathcal{E}_0(i, j) = \{\underline{1}_{M_z}^T \mathbf{X}_z(:, i) - \underline{\mu}^T \mathbf{X}_p(:, i) - \lambda_1(i) \geq \underline{1}_{M_z}^T \mathbf{X}_z(:, j) - \underline{\mu}^T \mathbf{X}_p(:, j)\}, \quad (40)$$

with the notation for S_z and \mathcal{S}_z as defined at the beginning of Section V. In the following discussion, since i is fixed, for notational simplicity we will use $\mathcal{E}_0(j) \triangleq \mathcal{E}_0(i, j)$. Note that, for **RoLpAl**, the error event is upper bounded by a similar expression as the above but with $\mathcal{E}_0(j)$ replaced by $\mathcal{E}_1(j) \triangleq \{\underline{1}_{M_z}^T \mathbf{X}_z(:, i) \geq \underline{1}_{M_z}^T \mathbf{X}_z(:, j)\}$. In order to analytically compare the performances of **RoLpAl** and **RoLpAl++**, we try to relate the events $\mathcal{E}_0(j)$ and $\mathcal{E}_1(j)$. Note that if $\mathcal{E}_0(j) \subseteq \mathcal{E}_1(j)$, then $\mathbb{P}(\mathcal{E}_0(j)) \leq \mathbb{P}(\mathcal{E}_1(j))$, and hence, **RoLpAl++** would outperform **RoLpAl**. Now, when $\underline{\mu} = \underline{0}_{M_z}$, $\mathcal{E}_0(j) \subseteq \mathcal{E}_1(j)$, $\forall j \in S_z$. For $\underline{\mu} \neq \underline{0}$, we divide the items in S_z into two disjoint groups:

- (a) $\lambda_2(j) > 0$: Since $\underline{\mu}^T \mathbf{X}_p(:, j) = 0$, $\underline{\mu}^T \mathbf{X}_p(:, i) \geq 0$ and $\lambda_1(i) \geq 0$, it follows that $\mathcal{E}_0 \subseteq \mathcal{E}_1$.
- (b) $\lambda_2(j) = 0$: We note that $\mathcal{E}_0(j) \subseteq \mathcal{E}_1(j) \cup \mathcal{E}'_1(j)$ where $\mathcal{E}'_1(j) = \{\underline{\mu}^T [\mathbf{X}_p(:, j) - \mathbf{X}_p(:, i)] \geq \kappa + \lambda_1(i)\}$, where $\kappa + \lambda_1(i) > 0$.

A technical problem, which does not allow us to state the categorical performance result, arises now. It is difficult to obtain the estimates for the dual variables $\underline{\mu}$ and hence of $\mathbb{P}(\mathcal{E}'_1(j))$. Therefore, we offer two intuitive arguments that provide insight into the relative performance of **RoLpAl++** and **RoLpAl**. The first argument is that the *majority* of the items in S_z will have $\lambda_2(j) > 0$ and thus, for a majority items in S_z , it follows that $\mathbb{P}(\mathcal{E}_0(j)) \leq \mathbb{P}(\mathcal{E}_1(j))$. This is because the set $\{j : \lambda_2(j) = 0\}$ is given by,

$$\left\{ j : (\underline{1}_{M_z}^T \mathbf{X}_z(:, j) - \underline{\mu}^T \mathbf{X}_p(:, j)) = \max_{\{l : \lambda_1(l) = 0\}} (\underline{1}_{M_z}^T \mathbf{X}_z(:, l) - \underline{\mu}^T \mathbf{X}_p(:, l)) \right\}, \quad (41)$$

and, as the number of tests increase and the number of nonzero components of $\underline{\mu}$ increase, the probability that above equality holds becomes smaller and smaller. Furthermore, for a small number of items $j \in S_z$ with $\lambda_2(j) = 0$, it is reasonable to expect that $\mathbb{P}(\mathcal{E}'_1(j))$ will be small. This is because the probability that

a defective item is tested in a pool with positive outcome is higher than the probability that a non-defective item is tested in a pool with positive outcome. Thus, the expected value of $\underline{\mu}^T[\mathbf{X}(:,j) - \mathbf{X}(:,i)]$ will be negative for a non-negative $\underline{\mu}$ and, thus using concentration of measure arguments, we can expect $\mathbb{P}(\mathcal{E}'_1(j))$ to be small. Thus, we expect that **RoLpAl++** to perform similar (or even better) than **RoLpAl**.

G. Chernoff Bounds

Theorem 4. (Bernstein Inequality [30]) Let X_1, X_2, \dots, X_n be independent real valued random variables, and assume that $|X_i| < c$ with probability one. Let $X = \sum_{i=1}^n X_i$, $\mu = \mathbb{E}(X)$ and $\sigma = \text{Var}(X)$. Then, for any $\delta > 0$, the following hold:

$$\mathbb{P}(X > \mu + \delta) \leq \exp\left(-\frac{\delta^2}{2\sigma^2 + \frac{2}{3}c\delta}\right) \quad (42)$$

$$\mathbb{P}(X < \mu - \delta) \leq \exp\left(-\frac{\delta^2}{2\sigma^2 + \frac{2}{3}c\delta}\right) \quad (43)$$

ACKNOWLEDGMENT

The authors thank Anuva Kulkarni for several interesting discussions in the initial phase of this work.

REFERENCES

- [1] R. Dorfman, “The Detection of Defective Members of Large Populations,” *The Annals of Mathematical Statistics*, vol. 14, no. 4, Dec. 1943.
- [2] D. Du and F. Hwang, *Pooling designs and non-adaptive group testing: Important tools for DNA sequencing*, World Scientific, 2006.
- [3] M. Sobel and P. A. Groll, “Group testing to eliminate efficiently all defectives in a binomial sample,” *Bell System Technical Journal*, vol. 38, no. 5, pp. 1179–1252, 1959.
- [4] H. Q. Ngo and Ding-Zhu Du, “A survey on combinatorial group testing algorithms with applications to DNA library screening,” in *Discrete mathematical problems with medical applications*, vol. 55 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pp. 171–182. Amer. Math. Soc., 2000.
- [5] A. J. Macula and L. J. Popyack, “A group testing method for finding patterns in data,” *Discrete Appl. Math.*, vol. 144, no. 1-2, pp. 149–157, 2004.
- [6] Dmitry M. Malioutov and Kush R. Varshney, “Exact Rule Learning via Boolean Compressed Sensing,” in *Proceedings of the International Conference on Machine Learning (ICML 2013)*, 2013, pp. 775–783.
- [7] Sanjeeb Dash, Dmitry M. Malioutov, and Kush R. Varshney, “Screening for learning classification rules via boolean compressed sensing,” in *Proc. ICASSP*, 2014, pp. 3360–3364.

- [8] J. Wolf, “Born again group testing: Multiaccess communications,” *IEEE Trans. Inf. Theory*, vol. 31, no. 2, pp. 185–191, Mar 1985.
- [9] Graham Cormode and S. Muthukrishnan, “What’s hot and what’s not: Tracking most frequent items dynamically,” *ACM Trans. Database Syst.*, vol. 30, no. 1, pp. 249–278, Mar. 2005.
- [10] A. C. Gilbert, M. A. Iwen, and M. J. Strauss, “Group testing and sparse signal recovery,” in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, Oct. 2008, pp. 1059–1063.
- [11] W. Kautz and R. Singleton, “Nonrandom binary superimposed codes,” *Information Theory, IEEE Transactions on*, vol. 10, no. 4, pp. 363–377, 1964.
- [12] P. Erdos, P. Frankl, and Z. Furedi, “Families of finite sets in which no set is covered by the union of r others,” *Israel Journal of Mathematics*, vol. 51, no. 1-2, pp. 79–89, 1985.
- [13] M Ruzinkó, “On the upper bound of the size of the r -cover-free families,” *J. Comb. Theory, Ser. A*, vol. 66, no. 2, pp. 302–310, 1994.
- [14] A. G. Dyachkov and V. V. Rykov, “Bounds on the length of disjunctive codes,” *Problems of Information Transmission*, vol. 18, no. 3, pp. 7–13, 1982.
- [15] A. Sebo, “On two random search problems,” *Journal of Statistical Planning and Inference*, vol. 11, no. 1, pp. 23–31, Jan. 1985.
- [16] G. Atia and V. Saligrama, “Boolean compressed sensing and noisy group testing,” *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, 2012.
- [17] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, “Non-adaptive group testing: Explicit bounds and novel algorithms,” *CoRR*, vol. abs/1202.0206, 2012.
- [18] D. Cabric, S. M. Mishra, D. Willkomm, R. Brodersen, and A. Wolisz, “A cognitive radio approach for usage of virtual unlicensed spectrum,” in *Proc. of 14th IST Mobile Wireless Communications Summit*, 2005.
- [19] FCC, “Et docket no. 02-155,” *Spectrum policy task force report*, Nov. 2002.
- [20] A. Sharma and C. R. Murthy, “Group testing based spectrum hole search for cognitive radios (accepted),” *IEEE Trans. Veh. Technol.*, 2014.
- [21] A. Sharma and C. R. Murthy, “On finding a subset of healthy individuals from a large population,” *CoRR*, vol. abs/1307.8240, 2013.
- [22] Hong-Bin Chen and F. K Hwang, “A survey on nonadaptive group testing algorithms through the angle of decoding,” *Journal of Combinatorial Optimization*, vol. 15, no. 1, pp. 49–59, 2008.
- [23] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, “Group testing with probabilistic tests: Theory, design and application,” *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 7057–7067, Oct. 2011.
- [24] M. Aldridge, L. Baldassini, and O. Johnson, “Group testing algorithms: Bounds and simulations,” *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3671–3687, June 2014.
- [25] J. Yoo, Y. Xie, A. Harms, W. U. Bajwa, and R. A. Calderbank, “Finding zeros: Greedy detection of holes,” *CoRR*, vol. abs/1303.2048, 2013.

- [26] D. M. Malioutov and M. B. Malyutov, “Boolean compressed sensing: L_p relaxation for group testing,” in *Proc. ICASSP*, 2012, pp. 3305–3308, IEEE.
- [27] M. B. Malyutov, “The separating property of random matrices,” *Mat. Zametki*, vol. 23, no. 1, pp. 155–167, 1978.
- [28] S. Aeron, V. Saligrama, and M. Zhao, “Information theoretic bounds for compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5111–5130, Oct. 2010.
- [29] Yurii Nesterov, *Introductory lectures on convex optimization : A basic course*, Kluwer Academic Publ., 2004.
- [30] G. Lugosi, *Concentration-of-Measure Inequalities*, Lecture Notes. Available at <http://www.econ.upf.edu/~lugosi/anu.pdf>, 2006.
- [31] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Mar. 2004.
- [32] D. Sejdinovic and O. Johnson, “Note on noisy group testing: Asymptotic bounds and belief propagation reconstruction,” in *Proc. Allerton Conf. on Commun., Control and Comput.*, 2010, pp. 998–1003.